Event Batching for ACTS

Charles Leggett

HEP-CCE PPS Meeting
2021/03/19

Concurrent Jobs

- Use nvidia-cuda-mps-server to do time-slicing of GPU
  - if can show a benefit, know there's room for improvement in the code
  - the opposite is not necessarily true

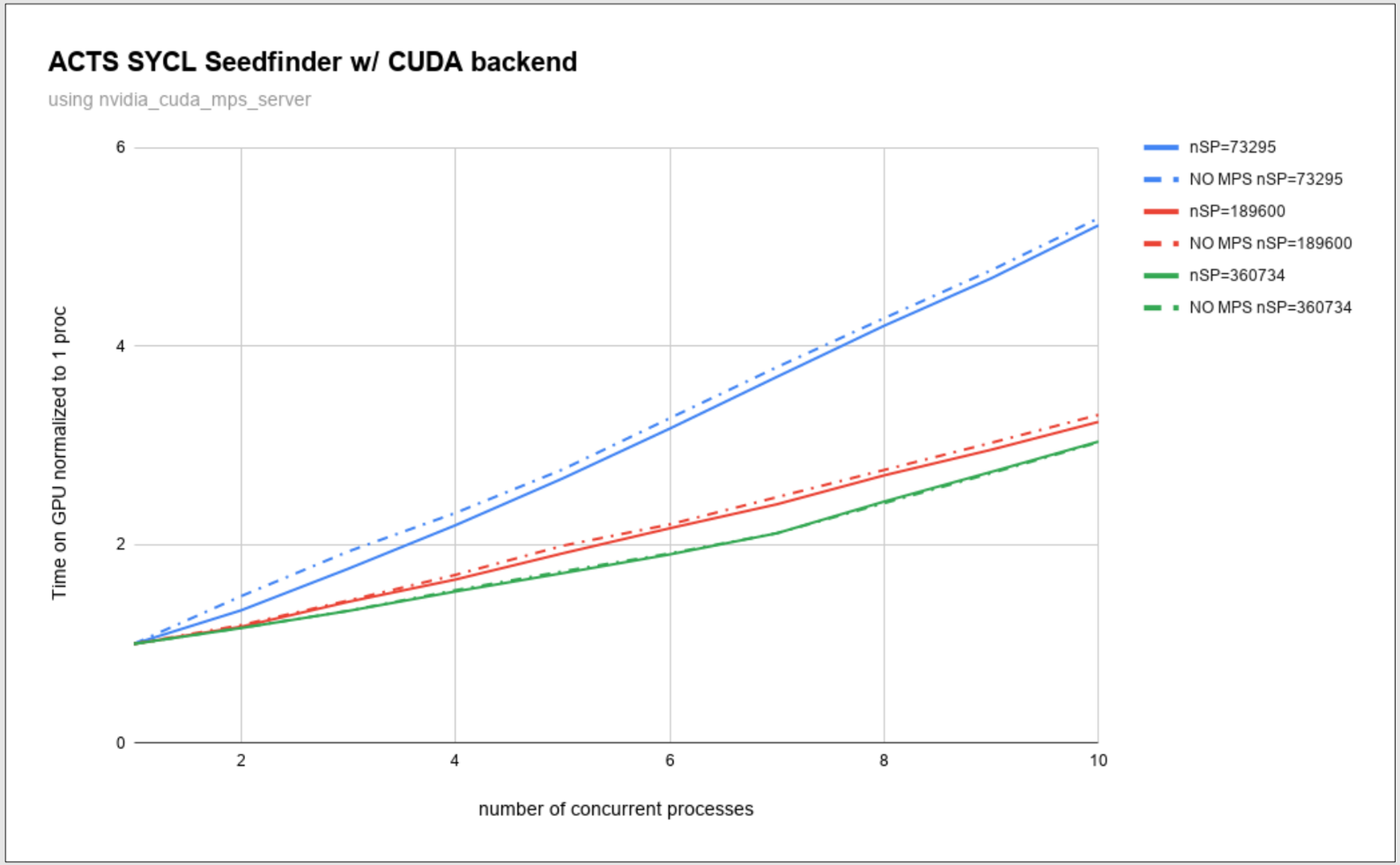
- ACTS SYCLSeedfinder w/ CUDA backend
  - memory manager is not thread safe
  - does a LOT of data reduction and pointers to indices on other arrays which would make it more challenging to add an extra index for "event number"

SYCL Seedfinder w/ CUDA Backend

- no obvious benefit to doing GPU timeslicing w/ mps-server
  - looking at the profile shows lots of GPU usage, mostly memory movement
  - not sure if it's the fault of the algorithm or related to sycl→cuda backend

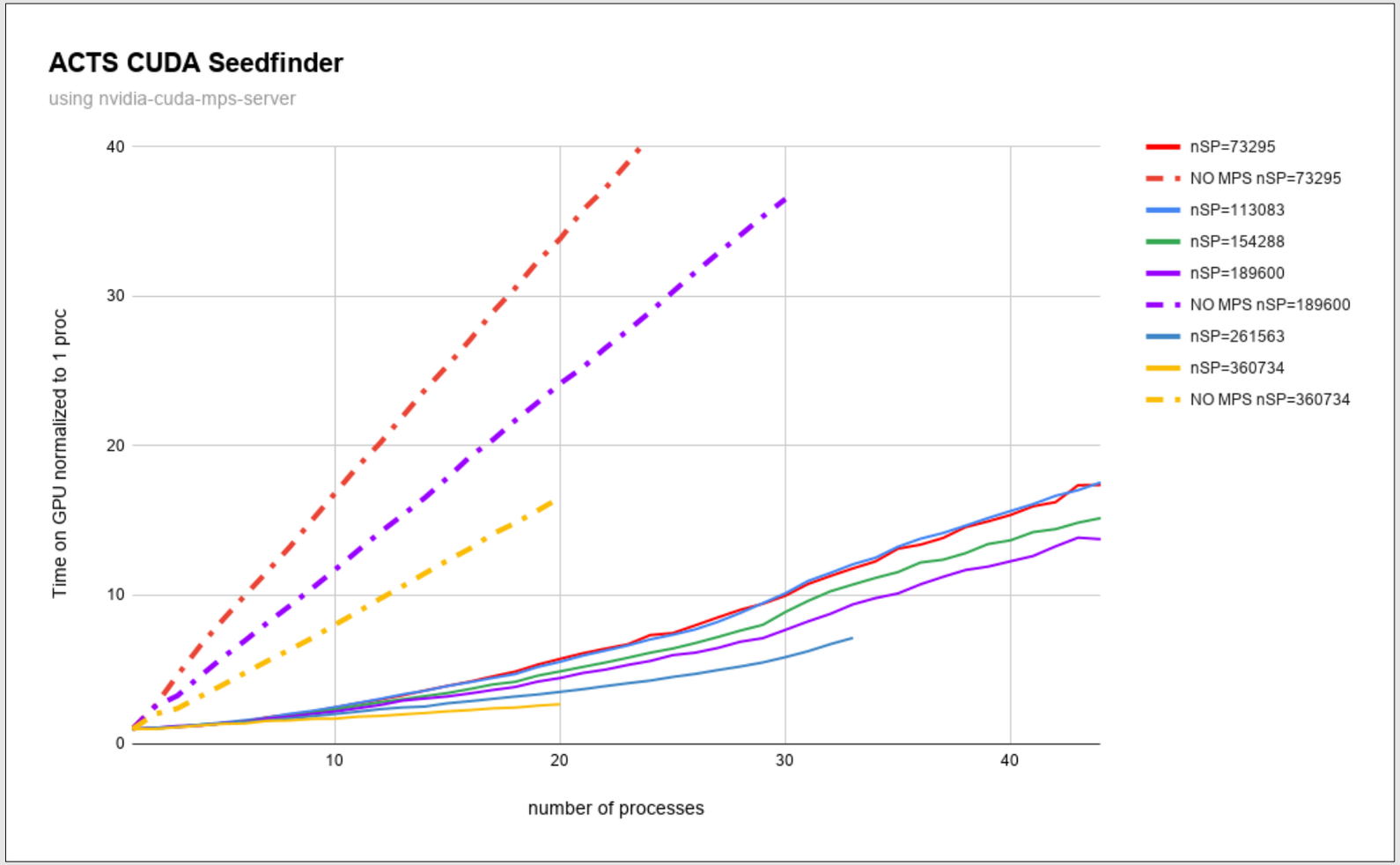
  - ran with 1-10 concurrent processes, 70k to 360k spacepoints

  - vertical axis is avg runtime for n concurrent processes, divided by runtime for 1 proccess



CUDA Seedfinder

- Much more promising
  - adding more processes results in small impact on total runtime
  - limited by amount of memory on device
- Code is not explicitly thread hostile ;-)
  - ACTS “framework” is (should be) thread safe
- May be more beneficial to process separate events in separate threads w/ different CUDA streams, allowing more mem/kernel interleaving
- Maybe do both to be sure…



fin